Spatially-Adaptive Feature Modulation for Efficient Image Super-Resolution - Supplemental Material -

Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan*
School of Computer Science and Engineering, Nanjing University of Science and Technology

Overview

In this document, we further demonstrate the effectiveness of the proposed spatially-adaptive feature modulation and the LayerNorm layer in Section 1. Then, we evaluate our method with the challenge winners in Section 2. We further compare the proposed method with ViT-based lightweight SR models and classical performance-oriented SR methods in Section 3 and Section 4, respectively. Next, we make some notes on the Urban100 dataset in Section 5. Finally, we show more visual comparisons in Section 6.

1. Ablations of the spatially-adaptive feature modulation and the LayerNorm

Effectiveness of the spatially-adaptive feature modulation. As described in the main paper, the proposed spatially-adaptive feature modulation layer consists of three components: feature modulation (FM), multi-scale representation (MR), and feature aggregation (FA). To intuitively illustrate what the SAFM layer learns, we show some learned features in Figure 1, where the corresponding features are extracted before the upsampling layer. Figure 1 demonstrates that the deep features learned with SAFM layers contain much richer feature information and attend to more high-frequency details, facilitating the reconstruction of high-quality images.

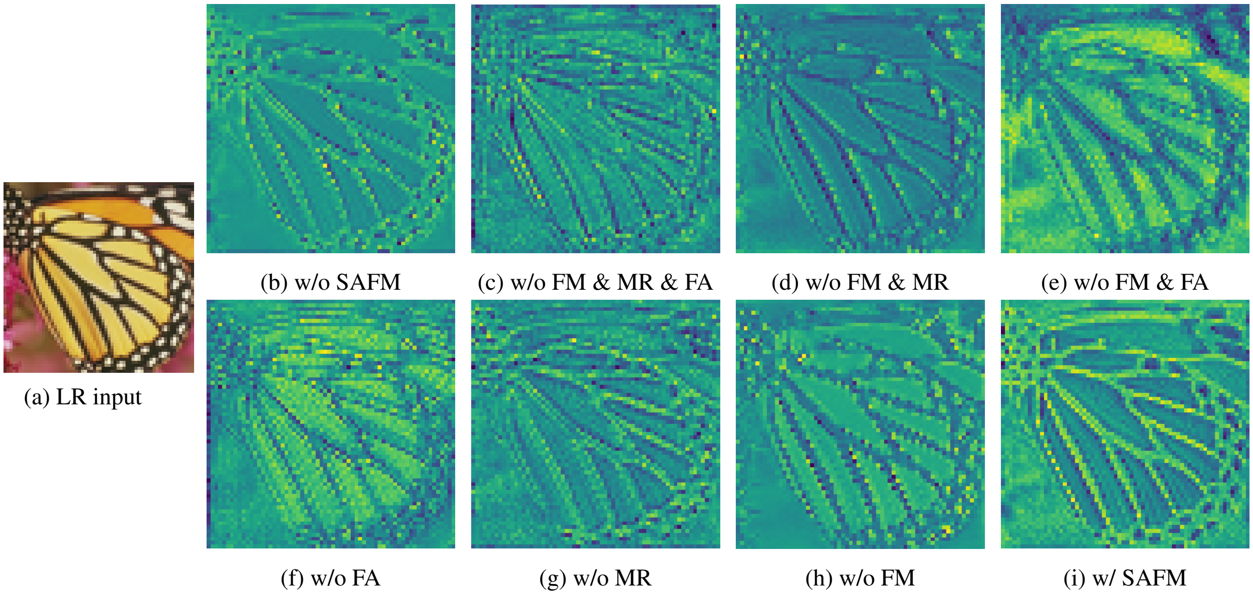
Effect of scales in the spatially-adaptive feature modulation. We evaluate the effect of features at different scales in the spatially-adaptive feature modulation (SAFM) layer on the $\times 4$ DIV2K validation set. Table 1 shows that removing any scale information deteriorates the reconstruction performance.

Effect of the LayerNorm layer. We show the visual results of different normalizations in Figure 2. As stated in the main paper, we obtained the results for the Frozen BatchNorm [5] and without the LayerNorm [2] before the training collapse. The model with BatchNorm layers generates images with unpleasant artifacts because it involves the estimated mean and variance of the entire training dataset during testing. The artifacts can be alleviated when we fix these estimates, as shown in Figure 2(c). Figure 2(d) and (f) show that applying normalization in the channel dimension can avoid the occurrence of artifacts. Compared to the L_2 normalization, the model with LayerNorm layers produces more precise results. We, therefore, introduce LayerNorm layers for stable training and well convergence. The reasons behind the LayerNorm remain to be further investigated.

2. Comparison with the challenge winners

We further compare our method with solutions of the challenge champion, i.e., RFDN (winner of AIM 2020 Efficient Super-Resolution Challenge [17]) and RLFN (winner of NTIRE 2022 Efficient Super-Resolution Challenge [10]). Table 2 demonstrates that our approach obtains a noticeable improvement in all measures except the running time. Table 5 of the main paper shows that our slower running time is mainly due to the use of LayerNorm [2], which requires the mean and standard deviation of the input features in the inference phase. Without LayerNorm, the runtime improves to 8.35ms, which is very close to the speed of RLFN. As described in Section 1, however, the importance of LayerNorm prevents us from removing this module directly. We will explore feasible alternatives in our future work.

*Corresponding author



(a) LR input

(b) w/o SAFM

(c) w/o FM & MR & FA

(d) w/o FM & MR

(e) w/o FM & FA

(f) w/o FA

(g) w/o MR

(h) w/o FM

(i) w/ SAFM

Figure 1. **Illustration of the learned deep features from ablations in the SAFM.** We average the features before the upsampling layer in the channel dimension and show the corresponding results. The proposed SAFM layer includes three components: feature modulation (FM), multi-scale representation (MR), and feature aggregation (FA). (h) indicates that the model pays less attention to the high-frequency regions without the feature modulation. (g) shows that the model fails to capture long-range information without the multi-scale representation. (f) illustrates the necessity of aggregating multi-scale features. The comparison of (i) with (b)-(h) suggests that the proposed method with the SAFM layer yields a finer feature representation with clearer structures that pays more attention to high-frequency details.

Table 1. **Effect of scales in the SAFM.** We evaluate the effect of features at different scales in the SAFM layer on the ×4 DIV2K validation set. The results show that removing any scale information affects the reconstruction performance.

Variants	SAFMN	w/o Scale 8	w/o Scale 8&4	w/o Scale 8&4&2
DIV2K_val	30.43/0.8372	30.39/0.8362	30.37/0.8357	30.34/0.8350

Table 2. **Efficiency comparison with the challenge winners on ×4 SR.** #GPU Mem. and #Avg. Time denote the maximum GPU memory consumption and the average running time of the inference phase, respectively. #FLOPs, #Acts and #Avg. Time are computed on an LR image with a resolution of 320 × 180 pixels. Our SAFMN obtains comparable performance and a better trade-off between reconstruction performance and model complexity.

Methods	#Params [K]	#FLOPs [G]	#Acts [M]	#GPU Mem. [M]	#Avg.Time [ms]	B100 [PSNR/SSIM]
RFDN [13]	433.45	23.82	98.46	176.75	7.23	27.60/0.7368
RLFN [7]	543.74	29.88	111.17	145.69	7.35	27.60/0.7364
SAFMN (Ours)	239.52	13.56	76.70	65.26	10.71	27.58/0.7359

3. Comparison with ViT-based lightweight SR methods

We compare the ×4 SR performance with ViT-based methods including ESRT [14], SwinIR-light [11], and ELAN-light [18]. We calculate their efficiency metrics in officially released codes with the fvcore library under super-resolving inputs to 1280 × 720 pixles. As these ViT-based lightweight SR methods have parameter sizes over 600K, we scale up the proposed SAFMN with 48 channels and 12 FMMs to 610K for a fair comparison. Table 3 shows that our SAFMN-c48n12 produces competitive results with much lower computational complexity. Compared to SwinIR-light, our method has 316.53K fewer parameters and is nearly 7× faster.

4. Comparison with classical SR models

To verify the scalability of SAFMN, we further compare the large version of SAFMN, which has 16 FMMs with 128 channels, with the state-of-the-art classical SR methods, including EDSR [12], RCAN [19], SAN [3], HAN [15], SwinIR [11]. Table 4 shows that our SAFMN shows significant advantages in terms of model efficiency compared to the evaluated CNN-based methods and obtains competitive reconstruction performances on five public benchmarks, benefiting from its capability of multi-scale feature modulation.

Table 3. **Comparison with ViT-based lightweight SR methods.** Our SAFMN-c48n12 produces competitive results with much lower computational complexity.

Methods	#Params [K]	#FLOPs [G]	#Acts [G]	#GPU Mem. [M]	#Avg.Time [ms]	Set14/Manga109 [PSNR]
ESRT [14]	751.77	298.32	6.92	6747.72	115.09	28.69/30.75
SwinIR-light [11]	929.63	61.69	1.28	368.19	130.28	28.77/30.92
ELAN-light [18]	640.39	54.12	1.09	240.40	41.70	28.78/30.92
SAFMN (Ours)	239.52	13.56	0.077	65.26	10.71	28.60/30.43
SAFMN-c48n12 (Ours)	613.10	34.84	0.149	90.14	16.61	28.77/30.93

Table 4. **Classical image SR results.** #Params and #FLOPs are measured under the setting of upscaling SR images to 1280 × 720 pixels on all listed scales. The proposed SAFMN achieves comparable performances with significantly less computational and memory costs.

Scale	Methods	#Params [M]	#FLOPs [G]	Set5	Set14	B100	Urban100	Manga109
×2	EDSR [12]	40.73	9387	38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351	39.10/0.9773
	RCAN [19]	15.45	3530	38.27/0.9614	34.12/0.9216	32.41/0.9027	33.34/0.9384	39.44/0.9786
	SAN [3]	15.86	3050	38.31/0.9620	34.07/0.9213	32.42/0.9028	33.10/0.9370	39.32/0.9792
	HAN [15]	63.61	14551	38.27/0.9614	34.16/0.9217	32.41/0.9027	33.35/0.9385	39.46/0.9785
	SAFMN (Ours)	5.56	1274	38.28/0.9616	34.14/0.9220	32.39/0.9024	33.06/0.9366	39.56/0.9790
	SwinIR [11]	11.75	2952	38.42/0.9623	34.46/0.9250	32.53/0.9041	33.81/0.9427	39.92/0.9797
×3	EDSR [12]	43.68	4470	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	34.17/0.9476
	RCAN [19]	15.63	1586	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	34.17/0.9476
	SAN [3]	15.90	1620	34.75/0.9300	30.59/0.8476	29.33/0.8112	28.93/0.8671	34.30/0.9494
	HAN [15]	64.35	6534	34.75/0.9299	30.67/0.8483	29.32/0.8110	29.10/0.8705	34.48/0.9500
	SAFMN (Ours)	5.58	569	34.80/0.9301	30.68/0.8485	29.34/0.8110	28.99/0.8679	34.66/0.9504
	SwinIR [11]	11.94	1363	34.97/0.9318	30.93/0.8534	29.46/0.8145	29.75/0.8826	35.12/0.9537
×4	EDSR [12]	43.90	2895	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	31.02/0.9148
	RCAN [19]	15.59	918	32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087	31.22/0.9173
	SAN [3]	15.86	937	32.64/0.9003	28.92/0.7888	27.78/0.7436	26.79/0.8068	31.18/0.9169
	HAN [15]	64.20	3776	32.64/0.9002	28.90/0.7890	27.80/0.7442	26.85/0.8094	31.42/0.9177
	SAFMN (Ours)	5.60	321	32.65/0.9005	28.96/0.7898	27.82/0.7440	26.81/0.8058	31.59/0.9192
	SwinIR [11]	11.90	774	32.92/0.9044	29.09/0.7950	27.92/0.7489	27.45/0.8254	32.03/0.9260

Table 5. **Quantitative comparison results on the Urban100 dataset.** Our proposed method performs less well in PSNR/SSIM but is comparable to IMDN and LAPAR in perceptual metrics, including NIQE and LPIPS.

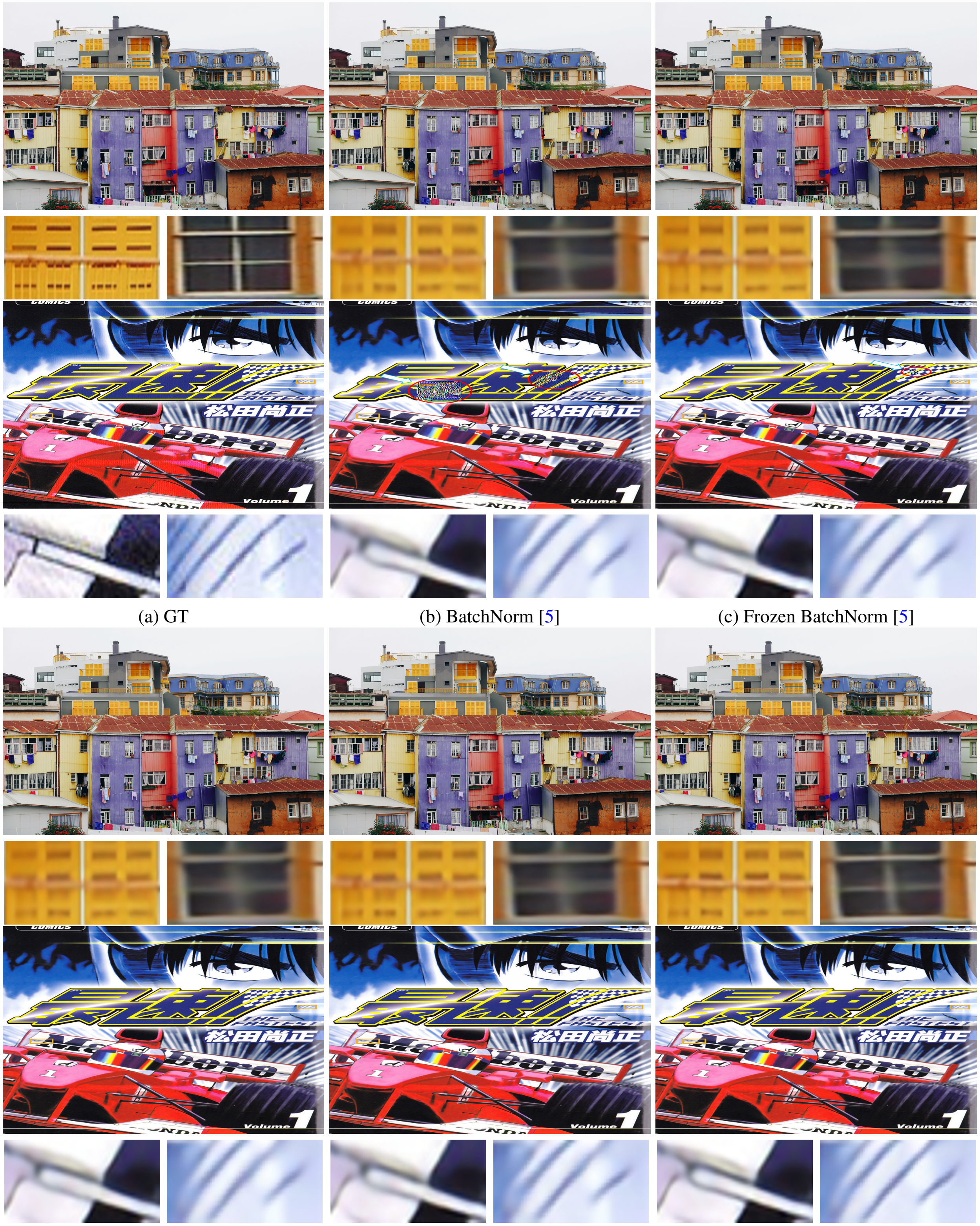
Scale	Methods	#Params [K]	#FLOPs [G]	#Acts [M]	PSNR	SSIM	NIQE	LPIPS
×2	IMDN [4]	694	159	423	32.17	0.9283	4.59	0.1132
	LAPAR-A [9]	548	171	677	32.17	0.9250	4.55	0.1129
	SAFMN (Ours)	228	52	299	31.84	0.9256	4.60	0.1138
×3	IMDN [4]	703	72	190	28.17	0.8519	5.21	0.2136
	LAPAR-A [9]	594	114	505	28.15	0.8523	5.21	0.2163
	SAFMN (Ours)	233	23	134	27.95	0.8474	5.28	0.2134
×4	IMDN [4]	715	41	108	26.04	0.7838	5.69	0.2879
	LAPAR-A [9]	659	94	452	26.14	0.7871	5.63	0.2868
	SAFMN (Ours)	240	14	77	25.97	0.7809	5.79	0.2881

5. Some notes on the Urban100 dataset

As shown in Table 5, the proposed SAFMN obtains a weak PSNR performance on the Urban100 dataset compared to other state-of-the-art methods, e.g., IMDN [4] and LAPAR-A [9]. The slight local luminance differences are responsible for these results. Since PSNR measures pixel-level distances rather than overall structure, slight differences in the luminance channel could lead to significant differences in PSNR. Furthermore, we visually compare images with a significant PSNR gap between our SAFMN and IMDN and observe no detectable changes in perceptual quality. Thus, we reevaluate these results using two commonly-used perceptual metrics: NIQE and LPIPS. Table 5 lists the quantitative results, and the proposed method achieves comparable performance to IMDN and LAPAR-A in terms of NIQE and LPIPS.

6. More visual results

In this section, we present additional visual comparisons with state-of-the-art methods [6, 1, 8, 4, 16] on the ×4 Urban100 dataset. Figure 3 shows that the proposed algorithm generates clearer images with finer detailed structures than those by state-of-the-art methods.



(a) GT

(b) BatchNorm [5]

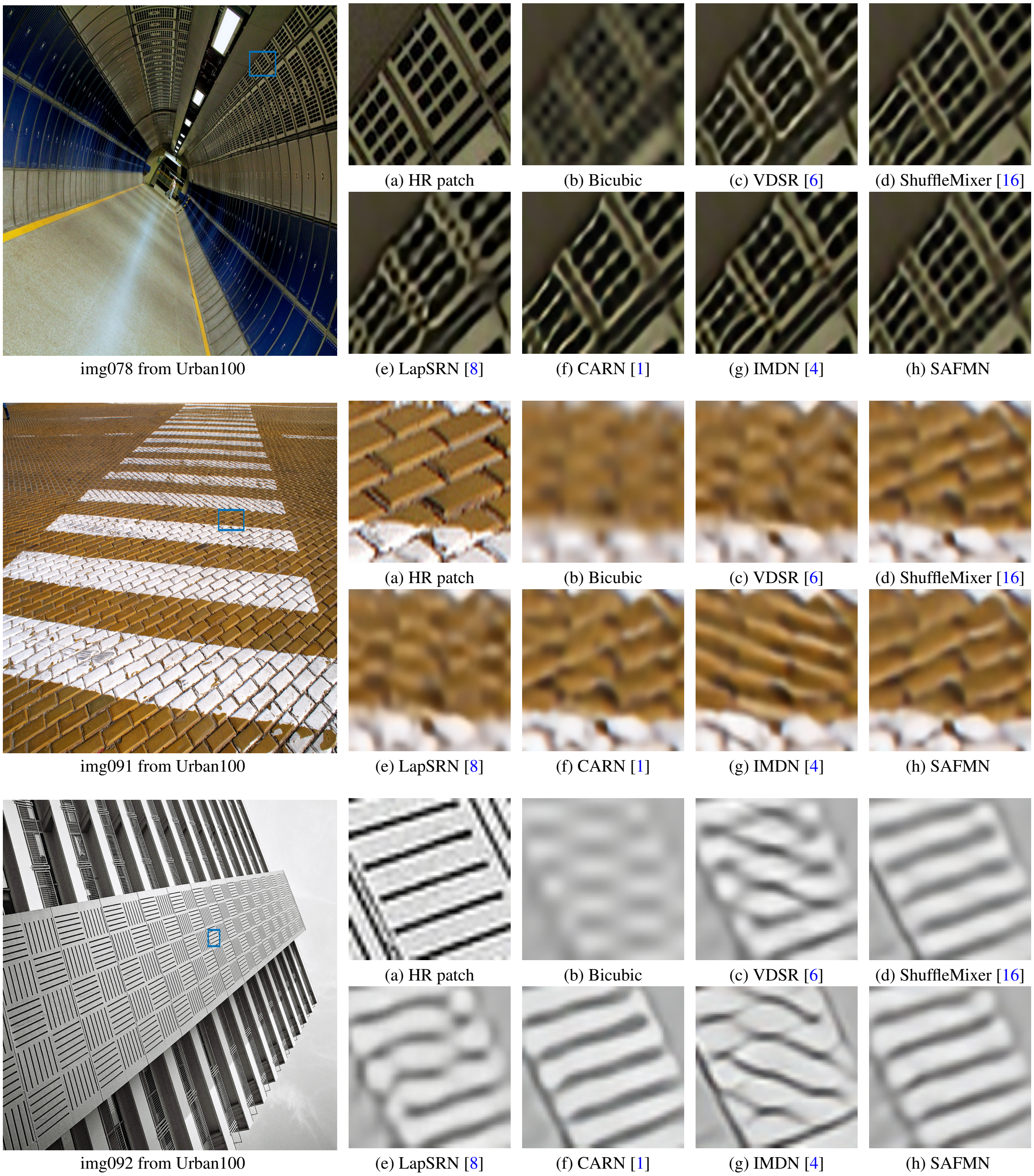
(c) Frozen BatchNorm [5]

(d) L_2 normalization

(e) w/o LayerNorm [2]

(f) w/ LayerNorm [2]

Figure 2. **Visual results of different normalization methods.** The proposed model with LayerNorm layers reconstructs better images.



img078 from Urban100

(a) HR patch (b) Bicubic (c) VDSR [6] (d) ShuffleMixer [16]

(e) LapSRN [8] (f) CARN [1] (g) IMDN [4] (h) SAFMN

img091 from Urban100

(a) HR patch (b) Bicubic (c) VDSR [6] (d) ShuffleMixer [16]

(e) LapSRN [8] (f) CARN [1] (g) IMDN [4] (h) SAFMN

img092 from Urban100

(a) HR patch (b) Bicubic (c) VDSR [6] (d) ShuffleMixer [16]

(e) LapSRN [8] (f) CARN [1] (g) IMDN [4] (h) SAFMN

Figure 3. **Visual comparisons for ×4 SR on the Urban100 dataset.** Our method generates images with clearer structures.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018. 3, 5

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1, 4

[3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 2, 3

[4] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, 2019. 3, 5

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 1, 4

[6] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 3, 5

[7] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPR Workshops*, 2022. 2

[8] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 3, 5

[9] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In *NeurIPS*, 2020. 3

[10] Yawei Li, Kai Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2022 challenge on efficient super-resolution: Methods and results. In *CVPR Workshops*, 2022. 1

[11] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 2, 3

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017. 2, 3

[13] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCV Workshops*, 2020. 2

[14] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *CVPR Workshops*, 2022. 2, 3

[15] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 2, 3

[16] Long Sun, Jinshan Pan, and Jinhui Tang. ShuffleMixer: An efficient convnet for image super-resolution. In *NeurIPS*, 2022. 3, 5

[17] Kai Zhang, Martin Danelljan, Yawei Li, and et al. AIM 2020 challenge on efficient super-resolution: Methods and results. In *ECCV Workshops*, 2020. 1

[18] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, 2022. 2, 3

[19] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2, 3